D648–D654 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 30 November 2012*
doi:10.1093/nar/gks1202

metaMicrobesOnline: phylogenomic analysis of microbial communities

Dylan Chivian[1,*], Paramvir S. Dehal[1], Keith Keller[1] and Adam P. Arkin[1,2,*]

[1]Division of Physical Biosciences, Lawrence Berkeley National Laboratory and [2]Department of Bioengineering, University of California, Berkeley, CA 94720, USA

Received October 5, 2012; Accepted October 30, 2012

ABSTRACT

The metaMicrobesOnline database (freely available at http://meta.MicrobesOnline.org) offers phylogenetic analysis of genes from microbial genomes and metagenomes. Gene trees are constructed for canonical gene families such as COG and Pfam. Such gene trees allow for rapid homologue analysis and subfamily comparison of genes from multiple metagenomes and comparisons with genes from microbial isolates. Additionally, the genome browser permits genome context comparisons, which may be used to determine the closest sequenced genome or suggest functionally associated genes. Lastly, the domain browser permits rapid comparison of protein domain organization within genes of interest from metagenomes and complete microbial genomes.

INTRODUCTION

Microbial community analysis using direct sequencing of DNA extracted from the environment, so-called 'environmental genomics' or 'metagenomics', is a rapidly changing field that is yielding an ever-growing depth of data and improved understanding of natural systems (1). The quantity of sequence one can obtain for the same cost is increasing exponentially (2); at the same time, longer regions of DNA are becoming available and therefore yielding more complete protein sequences at the individual sequence 'read' level. Additionally, improvements in approaches to 'binning' (3), that of grouping sequence reads into groupings that correspond to one or related strain 'phylotypes', as well as efforts to assemble data into the original longer sequence from the genome (4), the 'contigs', are offering the opportunity for beginning to be able to analyse larger contigs and even groups of contigs as putative 'draft' genomes extracted from metagenomic sequence (5). Additionally, in the near future there may be data sets that combine very long read technologies (6) or single-cell sequencing (7) with high-fidelity shorter read sequencing (8) for assembly of near complete microbial genomes without the need for culturing. Even today, there are experiments that have yielded complete and near-complete genomes directly from the environment (5,9,10). Although there are some powerful resources already in existence for metagenomic analysis, including MG-RAST (11), IMG/M (12) and CAMERA (13), additional approaches that take advantage of complete and near complete genomes to analyse the contigs and near full-length genes derived from metagenomes are needed, including phylogenomic resources. The metaMicrobesOnline database offers what we believe is the first phylogenetic gene tree resource that offers trees that include genes from both metagenomes and complete microbial genomes.

MATERIALS AND METHODS

The metaMicrobesOnline database extends the phylogenomic capabilities offered by MicrobesOnline (14) to include genes from metagenome assemblies. MetaMicrobesOnline does not perform contig assembly nor gene calling, focusing instead on gene tree analysis and leaving it to the user to determine the optimal approach for assembly and gene calling appropriate to their data. The public metagenomes that are currently available from metaMicrobesOnline have gene calls from IMG/M or MG-RAST, but any data set can be loaded as long as it reasonably conforms to an easily parsable format (e.g. FASTA for the contigs and tab-delimited gene coordinates that correspond to each contig). As full-length and near-full-length genes provide more reliable placement in gene trees, we have limited our analysis of the public metagenomes to those with longer

*To whom correspondence should be addressed. Tel: +1 510 495 8266; Fax: +1 510 495 2966; Email: DCChivian@lbl.gov
Correspondence may also be addressed to Adam P. Arkin. Tel: +1 510 495 2366; Fax: +1 510 486 6219; Email: APArkin@lbl.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

contigs that are likely to contain full-length genes (typically above about 500 bp to fit a single gene that is only a small domain, requiring contigs of 1000 bp and up to consistently obtain regular sized genes without truncation and longer for multi-domain proteins). Regrettably, the incomplete sequencing of even modestly complex microbial communities combined with the short read lengths of the current industry standard technologies and the need for advances in experimental design and assembly algorithms limits the number of metagenomes that are amenable to phylogenomic analysis. We expect as samples are more deeply sequenced, sequencing reads become longer, and assembly approaches improve that the number of metagenomes that produce non-truncated genes will increase, making multi-gene contig analysis such as offered by metaMicrobesOnline the norm for metagenomics.

Analysis with metaMicrobesOnline begins with contig sequences and gene calls being loaded into the metaMicrobesOnline analysis pipeline, where they are translated into protein sequences and scanned using HMMER3 (15) against canonical gene and protein domain families such as COG (16), Pfam (17) and TIGRFAMs (18). Alignments from the HMMER3 search are used to add the metegenomic genes to the multiple sequence alignment for each gene family. These augmented multiple sequence alignments are then used to build phylogenetic trees for each gene/domain family using FastTree-2 (19). It is possible to build trees even for gene families with hundreds of thousands of members because of the reduction in computational complexity that FastTree-2 offers, with memory $O[N^{1.25}L]$ and time $O[\log(N)N^{1.25}L]$. Membership of a given gene in gene/domain families is stored, the order of the domains within a gene and the order and orientation of genes in a contig. This information is available via interactive analysis tools such as the tree-based genome browser and the tree-based domain browser.

DATA AND TOOLS

Composition of the database

The metaMicrobesOnline database currently contains 1629 microbial isolate genomes (1429 bacterial genomes, 80 archaeal genomes and 120 eukaryotic fungal and algal genomes) and 155 metagenomes (123 ecological and 32 organismal-associated metagenomes). Unfortunately, at this time neither categorical (e.g. "hot spring") nor continuous (e.g. biogeochemical measurements) metadata about the samples is captured or used in analysis or selection of data sets for investigation, other than to include it where possible in the sample name. The database currently contains 7 million genes from

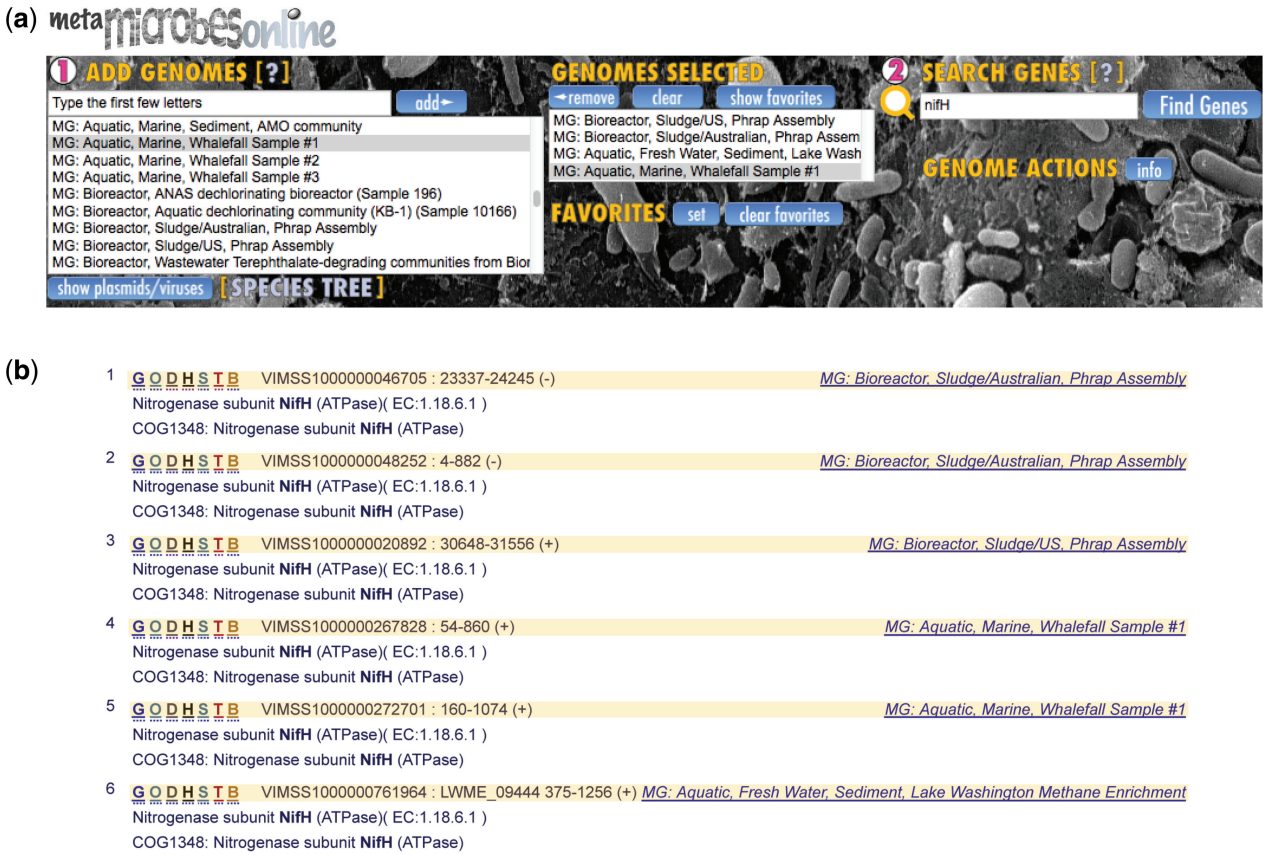

Figure 1. Selecting metagenomes and finding genes. (**a**) Genome and metagenome selector. Metagenome data sets identified with 'MG:' at the beginning. Name search for isolate genome or metagenome name in upper box, or scroll and click on desired data sets to add to selected set. Keyword search and genome information on right. (**b**) Results from keyword search for 'nifH' in several metagenomes (results truncated for clarity).

microbial isolates and 18 million genes from metagenomes contained in 4873 COG trees, 12148 Pfam trees and 3809 TIGRFAM trees. Among the largest trees are the PF00005 tree for 'ABC transporter ATPase subunit' (with 178 635 leaves), the PF07690 tree for the transporter 'major facilitator superfamily MFS-1' (with 112 515 leaves) and the PF00072 tree for 'signal transduction response regulator, receiver region' (with 99 438 leaves). COG and TIGRFAM typically detect fewer genes and therefore are usually smaller than their Pfam counterparts. Genes are not simply categorized as members of gene families or limited to lists of BLAST-based pairwise relationships (although such lists are available), but are rather placed phylogenetically into gene and domain families thus permitting a more wholistic view for functional inference.

Navigating to genes and tools

Analysis begins by selecting the metagenomes and genomes of interest using the 'genome/metagenome selector' (Figure 1a). The user can then continue onto genome/metagenome summary information (number of protein coding genes, overall COG functional category counts, etc.) using the 'genome info' button, or perform a targeted keyword search of the gene annotation information in the 'search field'. Acceptable search terms include canonical gene families (e.g. 'COG0001'), free text likely to occur in the description of those annotations (e.g. 'xylanase') or, if such names exist, additional gene names such as locus_tag or other synonyms for the gene. A list of genes that match the keyword is returned (Figure 1b) along with brief descriptions of the annotations and which metagenome or genome the gene is from. Quick links to information about the gene, such as gene summary ('G'), gene and domain family hits ('D') (Figure 2a), FastBLAST (20) determined homologues ('H') in microbial genomes and metagenomes (Figure 2b) and tree-based genome browser 'T' (Figure 3) are available from this view. The domains page (Figure 2a) also offers quick links to the tree-based genome browser ('T'), which includes the proximal genome context for related metagenomic contigs and genomes with ordering governed by the tree for the requested domain family. The domains page also offers a quick link to the tree-based domain browser ('D'), which shows the individual genes that possess the requested domain family and the other domains within those related genes (Figure 4).

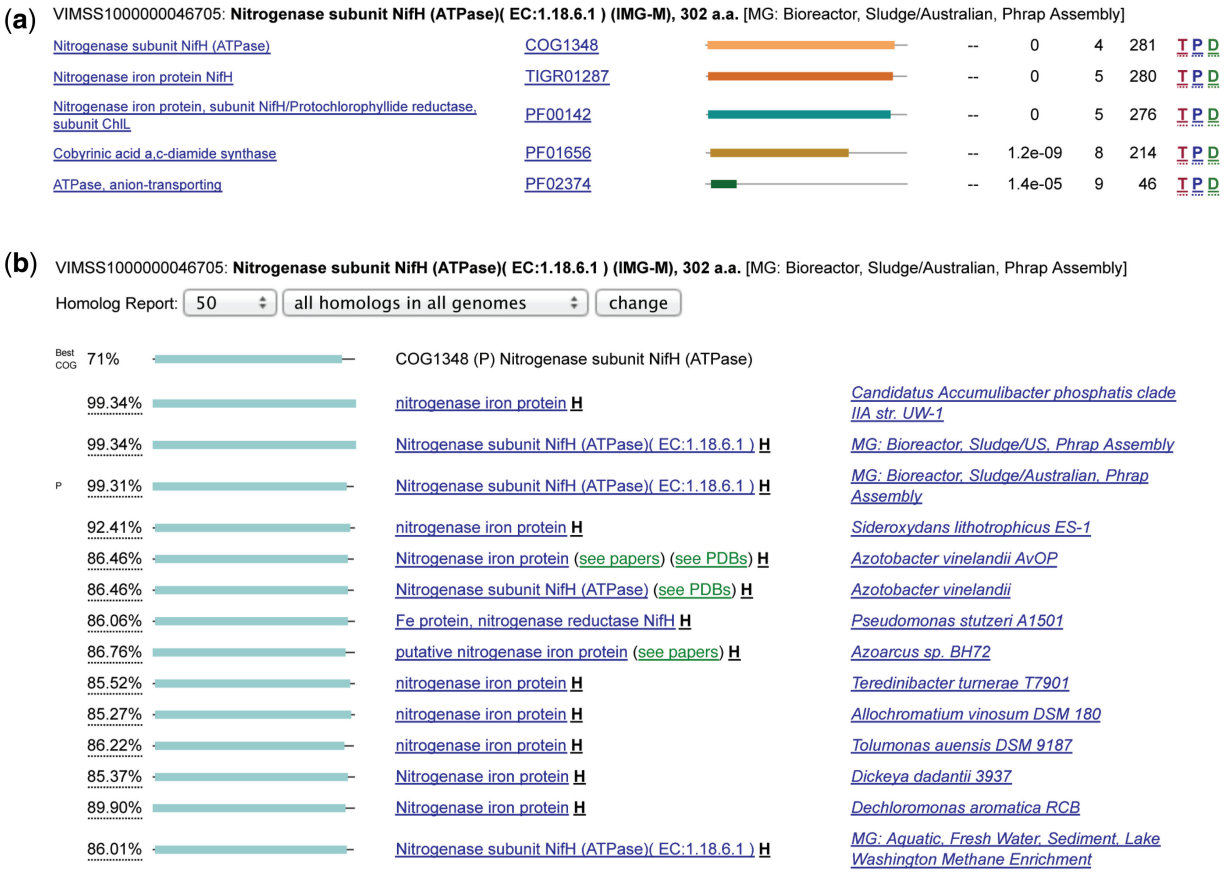

Figure 2. Gene and domain family page and Homologues. (**a**) Canonical gene and domain families, COG, Pfam and TIGRFAM assignments, including graphical depiction of region of gene matched to model, e-value of match, beginning and end position in gene of match and quick links to tree-based genome browser with tree based on the given gene/domain family ('T'), phylogenetic distribution of gene/domain family in microbial species tree ('P') and the tree-based domain browser ('D'). (**b**) Homologous genes found by FastBLAST in microbial genomes and metagenomes. Columns indicate duplicates in (meta)genome, with 'paralogs' indicated with 'P', graphical region of match in gene of interest, sequence identity of match, brief annotation of match (including links to papers and PDBs, if any) and the metagenome or species name where the gene is found. Clicking on graphical region of match shows pair alignment of match. Genes from metagenomes indicated by 'MG:' in the source name.

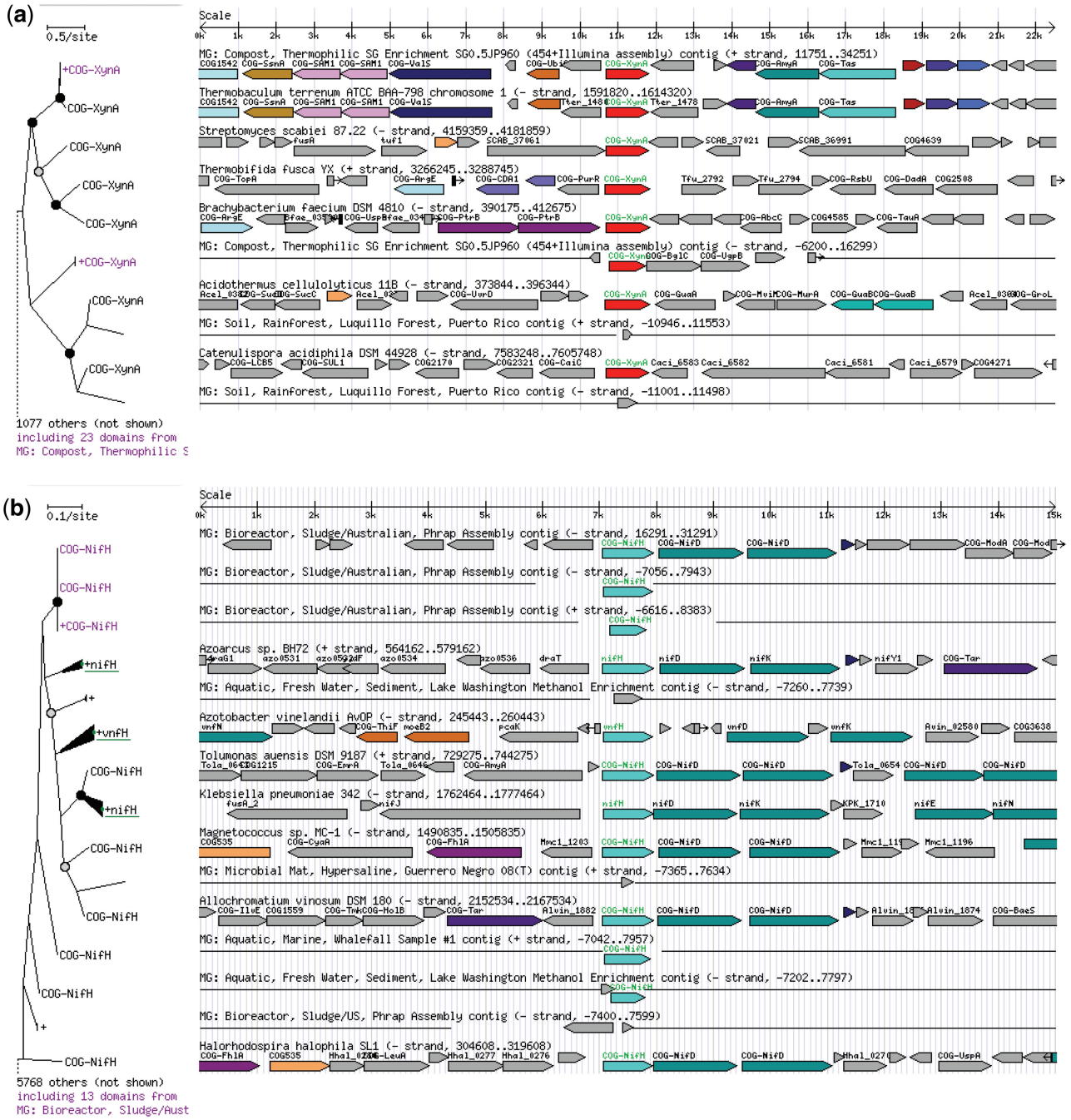

Figure 3. Tree-based genome browser. Local region of gene tree on left and local region of genome or contig on right (not shown: configuration of gene/domain family, percentage identity for collapsing closely related genes, number of rows to display, update button and zooming and panning options). Genes in same COG are shown in same colour. Any gene in browser can be clicked on to show more information or to reset as the target gene (if it has an assignment to a gene/domain family). Contigs shorter than window have lines indicating edge of truncation. (**a**) Strong synteny between compost metagenome contig and *Thermobaculum terrenum* genome increases confidence in species assignment. (**b**) NifH genes in metagenomes and microbial genomes show proximal conservation of related system genes, information that may be used for discovery of novel system components.

DISCUSSION

Using phylogeny and synteny to assign species

Determination of the gene complement of phylotypes within a community requires assignment of the genes and contigs to the source species. Although individual sequence binning approaches using nucleotide sequence signatures can suggest the taxonomic grouping, this is not always possible owing to the more rapid divergence of DNA compared with protein sequence. Further, taxonomic classification of genes and contigs using protein sequence [e.g. MEGAN (21)] suffers from the uncertainty

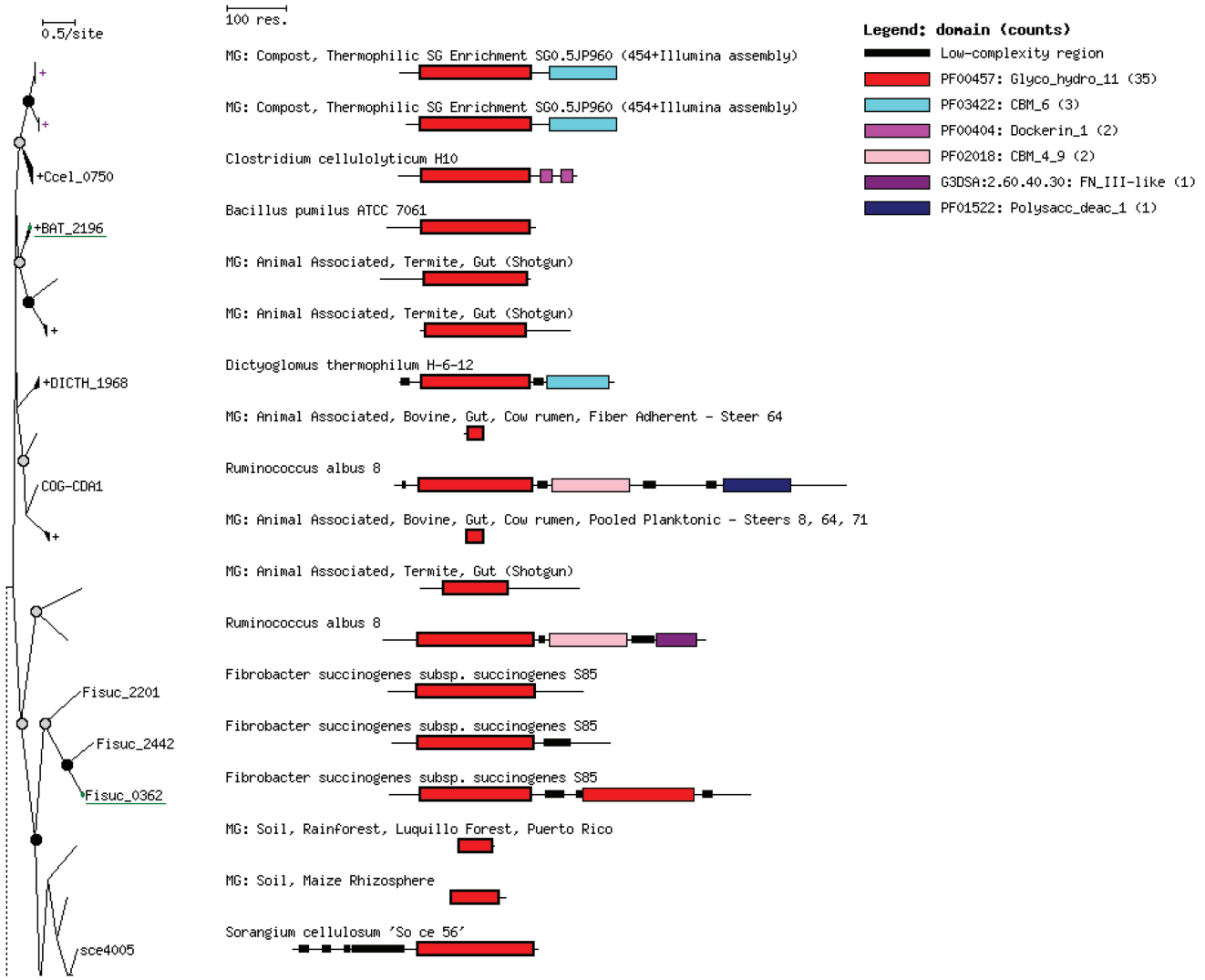

Figure 4. Tree-based domain browser. Local region of PF00457 (GH11) tree with genes from both metagenomes and microbial genomes. Domain region matched identified in red. Additional domains are shown in other colours. Image truncated for clarity.

presented by horizontal transfer of genes. Two approaches are available to mitigate these complications that take advantage of phylogenetic gene trees when one has multiple gene contigs available. Using the gene trees, one can identify the nearest neighbour species for the homologous genes for each gene in the contig and develop a consensus assignment for the contig as a whole. Second, certain gene families are more reliable for phylogenetic assignment, as they are not subject to horizontal transfer, as the presence of such 'housekeeping genes' (e.g. ribosomal proteins, rpoB, recA, etc.) from different species is detrimental to the new host. Identification of these genes within a contig can be used to more confidently assign the taxonomic grouping of the entire contig. Lastly, when a close relative with a sequenced genome is available, one can use the tree-based genome browser (Figure 3a) to examine genome context conservation to determine which species may be closest to the strain found in the metagenome.

Using genome and domain conservation and phylogeny to assign function

When species have significantly diverged and conservation of gene proximity is not merely indicative of a close relative, gene families that proximally co-occur across distantly related genomes often indicate genes that are members of a functional system (22). The tree-based genome browser can rapidly suggest which gene families should be investigated as part of a system (Figure 3b) and even suggest the function of the system if some of the co-occurring gene families have been characterized. Additionally, examination of the domain composition of the gene of interest using the tree-based domain browser (Figure 4) can reveal which domains are present in a truncated gene from a metagenome or show which domain combinations occur along with that domain family in other genes from metagenomes and isolate genomes. Lastly, phylogeny can be used to suggest function, as conserved function within the subfamily of

Figure 5. CdhA gene subfamily enrichment. Contigs from anaerobic methane-oxidizing community (AMO) are indicated with red asterisk. Lack of upstream synteny for very closely related carbon monoxide dehydrogenase genes with AMO community suggests expansion of gene by either horizontal transfer or lineage-specific expansion. Image truncated for clarity.

a gene family may be putatively propagated to the unknown gene.

Identifying environment-specific subfamily expansions

Horizontal transfer and lineage-specific expansions are two mechanisms by which additional copies of fitness-conferring genes are introduced to the gene pool (23). Phylogenetic gene trees can reveal which gene subfamilies are enriched within a given metagenome. This is especially useful when coarse gene family counting approaches suggest similar functional profiles when in fact different subfamilies of the gene tree, perhaps indicative of different functions such as different substrate specificities (24), may be preferentially enriched in one community over another. These gene trees, especially when coupled with taxonomic assignment or genome context, can reveal gene subfamily expansions that may be coupled with a fitness benefit in that given environment and serve as functional markers for a given ecosystem. For example, Figure 5 shows the expansion, as indicated by the relatively short branch lengths, of a subfamily of the carbon monoxide dehydrogenase gene in an anaerobic methane-oxidizing community (25). Genome context comparison of even very closely related cdhA genes shows no synteny upstream of the cdhA gene, indicating these are not merely duplicate contigs, and therefore this gene subfamily is considerably enriched, either by horizontal transfer, lineage-specific expansion or a mixture of these mechanisms.

CONCLUSIONS

Phylogenomic approaches to analysis of microbial communities that incorporate information from sequenced isolates and metagenomes permit both higher resolution functional comparisons between communities and enhance the ability to assign functions to species. The metaMicrobesOnline database makes such investigations possible with the use of interactive tools that permit rapid analysis and hypothesis generation.

ACKNOWLEDGEMENTS

The authors thank the JBEI researchers Steven Singer, Patrik D'haeseleer, Kristen DeAngelis, John Gladden,

Jean VanderGheynst, Janet Jansson, Michael Thelen, Blake Simmons and Terry Hazen for the provision of their data.

FUNDING

This work, performed by the Joint BioEnergy Institute (JBEI), was supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. [DE-AC02-05CH11231] between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. Funding for open access charge: U.S. Department of Energy.

Conflict of interest statement. None declared.

REFERENCES

1. DeLong,E.F. (2009) The microbial ocean from genomes to biomes. *Nature*, **459**, 200–206.
2. Mardis,E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.
3. Kunin,V., Copeland,A., Lapidus,A., Mavromatis,K. and Hugenholtz,P. (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, **72**, 557–578, Table of Contents.
4. Pop,M. (2009) Genome assembly reborn: recent computational challenges. *Brief. Bioinform.*, **10**, 354–366.
5. Wrighton,K.C., Thomas,B.C., Sharon,I., Miller,C.S., Castelle,C.J., VerBerkmoes,N.C., Wilkins,M.J., Hettich,R.L., Lipton,M.S., Williams,K.H. *et al.* (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, **337**, 1661–1665.
6. Maglia,G., Heron,A.J., Stoddart,D., Japrung,D. and Bayley,H. (2010) Analysis of single nucleic acid molecules with protein nanopores. *Methods Enzymol.*, **475**, 591–623.
7. Ishoey,T., Woyke,T., Stepanauskas,R., Novotny,M. and Lasken,R.S. (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.*, **11**, 198–204.
8. Bravo,H.C. and Irizarry,R.A. (2010) Model-based quality assessment and base-calling for second-generation sequencing data. *Biometrics*, **66**, 665–674.
9. Tyson,G.W., Chapman,J., Hugenholtz,P., Allen,E.E., Ram,R.J., Richardson,P.M., Solovyev,V.V., Rubin,E.M., Rokhsar,D.S. and Banfield,J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.
10. Chivian,D., Brodie,E.L., Alm,E.J., Culley,D.E., Dehal,P.S., DeSantis,T.Z., Gihring,T.M., Lapidus,A., Lin,L.H., Lowry,S.R. *et al.* (2008) Environmental genomics reveals a single-species ecosystem deep within Earth. *Science*, **322**, 275–278.
11. Glass,E.M., Wilkening,J., Wilke,A., Antonopoulos,D. and Meyer,F. (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb. Protoc.*, **2010**, pdb prot5368.
12. Markowitz,V.M., Chen,I.M., Chu,K., Szeto,E., Palaniappan,K., Grechkin,Y., Ratner,A., Jacob,B., Pati,A., Huntemann,M. *et al.* (2012) IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.*, **40**, D123–D129.
13. Sun,S., Chen,J., Li,W., Altintas,I., Lin,A., Peltier,S., Stocks,K., Allen,E.E., Ellisman,M., Grethe,J. *et al.* (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.*, **39**, D546–D551.
14. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
15. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
16. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
17. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
18. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
19. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
20. Price,M.N., Dehal,P.S. and Arkin,A.P. (2008) FastBLAST: homology relationships for millions of proteins. *PLoS One*, **3**, e3589.
21. Huson,D.H. and Mitra,S. (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol. Biol.*, **856**, 415–429.
22. Jacob,F., Perrin,D., Sanchez,C. and Monod,J. (1960) Operon: a group of genes with the expression coordinated by an operator [in French]. *C R Hebd. Seances Acad. Sci.*, **250**, 1727–1729.
23. David,L.A. and Alm,E.J. (2011) Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**, 93–96.
24. Chen,Z., Friedland,G.D., Pereira,J.H., Reveco,S.A., Chan,R., Park,J.I., Thelen,M.P., Adams,P.D., Arkin,A.P., Keasling,J.D. *et al.* (2012) Tracing determinants of dual substrate specificity in glycoside hydrolase family 5. *J. Biol. Chem.*, **287**, 25335–25343.
25. Hallam,S.J., Putnam,N., Preston,C.M., Detter,J.C., Rokhsar,D., Richardson,P.M. and DeLong,E.F. (2004) Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*, **305**, 1457–1462.